A recipe for creating ideal hybrid memristive-CMOS neuromorphic processing systems

Chicca, E ; Indiveri, G

Abstract: The development of memristive device technologies has reached a level of maturity to enable the design and fabrication of complex and large-scale hybrid memristive-Complementary Metal-Oxide Semiconductor (CMOS) neural processing systems. These systems offer promising solutions for implementing novel in-memory computing architectures for machine learning and data analysis problems. We argue that they are also ideal building blocks for integration in neuromorphic electronic circuits suitable for ultra-low power brain-inspired sensory processing systems, therefore leading to innovative solutions for always-on edge-computing and Internet-of-Things applications. Here, we present a recipe for creating such systems based on design strategies and computing principles inspired by those used in mammalian brains. We enumerate the specifications and properties of memristive devices required to support always-on learning in neuromorphic computing systems and to minimize their power consumption. Finally, we discuss in what cases such neuromorphic systems can complement conventional processing ones and highlight the importance of exploiting the physics of both the memristive devices and the CMOS circuits interfaced to them. We wish to acknowledge Melika Payvand and Regina Dittmann for the constructive comments on this manuscript. The illustration of Fig. 1 was kindly provided by the University of Zurich, Information Technology, MELS/SIVIC, Sarah Steinbacher. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program Grant Agreement No. 724295 (NeuroAgents).

DOI: https://doi.org/10.1063/1.5142089

A recipe for creating ideal hybrid memristive-CMOS neuromorphic computing systems

E. Chicca[1, a)] and G. Indiveri[2, b)]
[1)] *Faculty of Technology and Cognitive Interaction Technology - Center of Excellence (CITEC), Bielefeld University, Bielefeld, Germany*
[2)] *Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland*

(Dated: 13 December 2019)

The development of memristive device technologies has reached a level of maturity to enable the design of complex and large-scale hybrid memristive-CMOS neural processing systems. These systems offer promising solutions for implementing novel in-memory computing architectures for machine learning and data analysis problems. We argue that they are also ideal building blocks for the integration in neuromorphic electronic circuits suitable for ultra-low power brain-inspired sensory processing systems, therefore leading to the innovative solutions for always-on edge-computing and Internet-of-Things (IoT) applications. Here we present a recipe for creating such systems based on design strategies and computing principles inspired by those used in mammalian brains. We enumerate the specifications and properties of memristive devices required to support always-on learning in neuromorphic computing system and to minimize their power consumption. Finally, we discuss in what cases such neuromorphic systems can complement conventional processing ones and highlight the importance of exploiting the physics of both the memristive devices and of the CMOS circuits interfaced to them.

Neuromorphic computing has recently received considerable attention as a discipline that can offer promising technological solutions for implementing power- and size-efficient sensory-processing, learning, and Artificial Intelligence (AI) applications[1–5], especially in cases in which the computing system has to operate autonomously "at the edge", i.e., without having to connect to powerful (but power hungry) server farms in the "cloud". The term "neuromorphic" was originally coined in the early 90's by Carver Mead to refer to mixed signal analog/digital Very Large Scale Integration (VLSI) computing systems based on the organizing principles used by the biological nervous systems[6]. In that context, "neuromorphic engineering" emerged as an interdisciplinary research field deeply rooted in biology that focused on building electronic neural processing systems by exploiting the physics of silicon to directly "emulate" the bio-physics of real neurons and synapses. More recently the definition of the term "neuromorphic" has been extended in two additional directions: on one hand to describe more generic spike-based processing systems engineered to "simulate" spiking neural networks for the exploration of large-scale computational neuroscience models[7–9]; and on the other hand to describe dedicated electronic neural architectures that make use of both electronic Complementary Metal-Oxide Semiconductor (CMOS) circuits and memristive devices to implement neuron and synapse circuits[10,11].

Another recent and very promising trend in developing dedicated hardware architectures for building accelerated simulators of artificial neural networks is related to the field of machine learning and AI[12,13]. The types of neural networks being proposed within this context are only loosely inspired by biology, are aimed at high accuracy pattern recognition based on large data-sets, and require large amounts of memory for storing network states and parameters. While this approach is producing amazing results in a wide range of application areas, the computing systems used to simulate these networks use significant amount of compute resources and power, especially for the training phase: the learning algorithms rely on high precision digital representations for calculating high accuracy gradients, and they typically require the storage (and transfer from peripheral memory to central processing areas) of very large data-sets. Furthermore, they often separate the training from the inference phase, dismissing the ability to adapt to novel stimuli and changing environmental conditions, typical of biological systems.

FIG. 1. The ideal memristive neuromorphic computing system requires the right mix of CMOS circuits and memristive devices, as well as the proper use of spatial resources and temporal dynamics, that need to be well matched to the system's signal-processing applications and use-cases.

While there are examples of hybrid memristive-CMOS hardware architectures being developed to provide support for AI deep network accelerators[5,11,14,15], it is important to clarify that many of the hybrid memristive-CMOS neuromorphic

[a)] Electronic mail: chicca@cit-ec.uni-bielefeld.de
[b)] Electronic mail: giacomo@ini.uzh.ch

circuits proposed in the literature[16–20] as well as the original neuromorphic approach of emulating biological neural systems proposed by Mead, are distinct and complementary to the machine learning one. While the machine learning approach is based on software algorithms developed to minimize the recognition error in very specific pattern recognition tasks, the original neuromorphic approach is based on brain-inspired electronic circuits and hardware architectures designed for reproducing the function of cortical and biological neural circuits[21]. As a consequence, this approach aims at understanding how to build robust and low-power neural processing systems using inhomogeneous and highly variable components, fault-tolerant massively parallel arrays of computational elements, and in-memory computing (non von Neumann) information processing architectures[22]. In the following, when discussing about "hybrid CMOS-memristive neuromorphic computing systems", we will refer to this specific approach.

Our *recipe* (Fig. 1) for optimally building neuromorphic systems by co-integrating memristive devices with CMOS circuits is based on the following considerations.

a. Lay out the ingredients in parallel on the worktop. To minimize power consumption and maximize robustness to variability, it is important to use physically distinct instantiations of neuron and synapse circuits, distributed across the silicon substrate[23]. This strategy is very different from the one used to build classical computing systems based on the von Neumann architecture. In classical processors there is a single or a small number of computing blocks that are time-multiplexed at very high clock rates to execute calculations, or to simulate many "parallel" neural processes[7,9,24]. The continuous transfer of data between memory and the time-multiplexed processing unit(s) required to carry out computation is limited by the infamous von Neumann bottleneck[25], and is the major cause of high energy consumption. In contrast, the amazing energy efficiency of biological systems, and of the neuromorphic ones that emulate them, arises from the in-memory computing nature of their architectures: there are multiple instances of neuron and synapse elements that carry out computation and at the same time store the network state. The disadvantage of having distributed state-full neuron and synapse circuits is that it can require significant amount of silicon real-estate for integrating all their memory structures (e.g., see the 4.3 cm^2 IBM TrueNorth chip[24]). However, the progress in CMOS fabrication technologies, the emergence of monolithic 3D integration technologies, and the possibility to co-integrate nano-scale memristive devices with mixed-signal analog/digital CMOS circuits in advanced node processes can substantially mitigate this problem[26].

b. Take your time. By eliminating the need to use time-multiplexed processing elements, these neuromorphic processing architectures can be designed to run in real physical time (time represents itself) as it happens in real biological neural networks. This is a radical departure from the classical way of implementing computation, that has decoupled computer simulation time from physical time since the very early designs of both computing systems and artificial neural networks[27,28]. For sensory-motor processing systems and edge-computing applications that need to measure and process natural signals, this is a tremendous advantage. Allowing time to represent itself removes the need of complicated clock or synchronizing structures that would otherwise be required to track the passage of simulated time. All computing elements in such neuromorphic systems are then coupled through the common variable of real-time (e.g., for implementing binding by synchronization[29]). To build sensory-processing systems that are best tuned to the signals they are required to process (or that can learn to extract information from them), it is necessary to use neural processing and learning circuits that have the same time-constants and dynamics of their input signals (e.g., to create "matched-filter" that can naturally resonate with their inputs). In the case of natural signals typically processed by humans, such as voice or gestures, these time constants should range from milliseconds to minutes or longer. These time constants are extremely long, compared to the typical processing rates of digital circuits. This allows neuromorphic systems to reduce power consumption even more and to have very large bandwidths for seamlessly transmitting signals across the network and via I/O pathways in shared buses[30,31]. However, such long time constants can be very difficult to achieve using pure CMOS circuits[32]. Memristive devices offer an ideal solution to this limitation. Although such devices are usually treated as non-volatile memories, certain material systems exhibit a rather volatile resistance change after electrical biasing, with temporal scales that can be tuned and matched to biological neural and synaptic dynamics[33–35]. Recent demonstrations of volatile memristive devices used to model neural dynamics include the emulation of nociceptors (i.e., sensory neuron receptors able to detect noxious stimuli)[36] and the implementation of spike-timing dependent learning rules with tunable forgetting rates[37,38]. In addition to exploiting the physics of the memristive devices to tune their volatility properties, it is possible to co-design more complex hybrid memristive-CMOS neuromorphic circuits to implement the wide range of time constants needed to model the multiple plasticity phenomena observed in biology (ranging from milliseconds in synaptic short term depression to hours and more in structural plasticity) and crucial for artificial neural processing systems[32,39].

c. Don't worry about density. Memristive devices are often praised for the small (nano-scale) size, which can be exploited to develop very high density cross-bars[40] in which the memristive devices are used as a learning synapses[41]. Nevertheless, current high-density approaches are not able to produce learning dynamics sufficiently complex for solving real-world tasks (e.g., with matched temporal scales, or suitable for life-long learning requirements). The achievement of such dynamics in a single device requires sophisticated material engineering efforts which are still beyond the current state-of-the-art. Conversely, by dismissing the chimera of high density synaptic arrays and co-integrating nano-scale memory elements with mixed signal analog/digital neuromorphic circuits, it is possible to implement sophisticated learning mechanisms that can exploit many features of memristive devices, besides their compact footprint, such as non-volatility, stochasticity, or state-dependent conductance changes. Furthermore, combining multiple transistors with one or more

memristive devices enables the design of complex synapse circuits that can reduce the effect of variability[42], enable the control of stochastic switching behaviors[11,43,44], and produce linear or non-linear state-dependent weight-updates[45,46].

d. Play it by ear: variability and randomness. Memristive devices are affected by both device-to-device and cycle-to-cycle variability[47,48]. Significant material science and device technology research efforts are being made to minimize such variability[41,47,49–52]. However, rather than fighting these variability effects with different materials or device technologies, neuromorphic systems can be designed to embrace and exploit them[39]. Examples of theoretical neural processing frameworks that *require* variability can be found in the domain of ensemble learning[53], reservoir computing[54] and liquid state machines[55]. Current efforts in neuromorphic engineering for implementing such frameworks to solve spatio-temporal pattern recognition problems rely on the variability provided by transistor device-mismatch effects[56–60]. Integration of memristive devices with inhomogeneous properties in such architectures can provide a richer set of distributions useful for enhancing the computational abilities of these networks. Indeed, multiple circuit solutions have already been proposed to better control the shape and parameters of such distributions[11,42].

One important source of variability in the operational parameters of memristive devices is in their switching mechanism. In filamentary memristive devices, this mechanism exhibits stochastic behavior which stem from the underlying filament formation process[18,61–63]. This intrinsic probabilistic property of memristive devices can be exploited for implementing stochastic learning in neuromorphic architectures[43,44,48,64–66], which in turn can be used to implement faithful models of biological cortical microcircuits[67,68], solve memory capacity and classification problems in artificial neural network applications[69,70], and reduce the network sensitivity to their variability[43]. Recent results on stochastic learning modulated by regularization mechanisms, such as homeostasis or intrinsic plasticity[44,71–73], present an excellent potential for exploiting the features memristive devices, even when restricted to binary values.

e. Don't (hard) limit your devices. In the context of deploying always-on learning systems (both artificial and biological) in real-world applications, a critical feature is their memory storage capacity[74,75]. When designing hardware neuromorphic learning system that have practical physical restrictions or limitations on the available resources (such as the number of memory devices integrated in the system, their resolution, precision, or dynamic range) it is important to be aware of the *theoretical* limits that set the bounds of achievable memory capacity and learning performance, independent of the device properties.

The thorough theoretical analysis on the limits of memory capacity in neural processing systems presented by Fusi and Abbott in 2007[74] provides essential guiding principles for the construction of artificial learning memristive systems. In this analysis, learning models are subdivided into four main categories, according to two key features: the synaptic weight bounds (hard or soft) and the (im)balance of potentiation and depression. Hard bounds are limits on the synaptic weight values that cannot be exceeded. Soft bounds are limits that can only be reached in the asymptotic limit. Typically, in neural network models with hard bounds, the weight update step size is constant and therefore independent of the weight value itself. Conversely soft bounds are introduced by allowing weight updates to depend on synaptic strength and to decrease as they approach the bound itself.

Even though it is clear that in real physical systems hard bounds are unavoidable (e.g., the supply rails in an electronic system), there is evidence that memristive devices exhibit soft bounds[76]. Therefore, by combining CMOS circuits with memristive devices, it is possible to design hybrid circuits that can implement and control the devices soft bounds for improving learning at the network level and for improving the overall system performance, e.g., in terms of reduced power consumption and increased memory capacity[45]. In contrast, it is impossible to precisely balance positive changes of synaptic weights with negative ones in hybrid memristive-CMOS neuromorphic computing systems. Given this unbalanced potentiation and depression property, the longest memory lifetime is achieved thanks to soft bounds, independently of the specific model chosen among those investigated by Fusi and Abbott[74].

To best implement the recipe we proposed it is necessary to use the right list of *ingredients*: a combination of memristive devices with multiple complementary features. The recipe *shopping-list* should comprise devices with different properties on retention, endurance, variability, switching currents, on-off ratios, that can be interfaced to analog and digital electronic CMOS circuits. However, even before attempting to bake the final hardware neural processing system, it is important to have access to realistic and faithful device models, so that during the design phase it will be possible to specify the characteristics of both the CMOS and memristive components and understand how to best exploit their processing features for properly modeling the different aspects of plasticity and neural information processing systems.

Once fabricated, these neuromorphic processing systems should implement always-on life-long learning features so that they can adapt to changes in their input signals and keep a proper operating regime. This implies that the hybrid CMOS-memristive neuromorphic system would be updating its synaptic weights continuously, with every learning event. This requires the use of memristive devices that support small gradual conductance changes, and very small currents (e.g., $< 1\,\mu$A), to minimize power consumption. In this case, the retention rate of such devices does not need to be extremely long, but should be compatible with the rate of weight update (which can be seen as a "refresh" operation) in the system. For example, in typical "edge" sensory-processing applications (wearable devices, home automation, surveillance, environmental monitoring, etc.) this could range from milliseconds to seconds or minutes.

On the other hand, once the learning process has terminated or if there is a long pause in the rate of input signals (e.g., during the night in ambient monitoring tasks), then it will be useful to be able to consolidate the memories formed in non-volatile memristive devices with high on-off ratio and long-

retention rates. In this case, since this operation would not be as frequent as the weight-update one for the on-line learning case, it would be acceptable to use devices that require larger switching currents, and that have a small number (even two) stable states[77].

To match the time constants of the neural processing system to the dynamics of its input signals, to maintain a stable operating region over long time scales, and to optimize the learning of complex spatio-temporal patterns, it is necessary to implement both fast (short term depression, long term potentiation, long term depression, etc.) and slow (intrinsic, homeostatic, structural) plasticity mechanisms, "orchestrating" multiple time-scales in the learning circuits[78]. For this it is crucial to be able to use volatile memristive devices that span a wide range of retention rates (e.g., from milliseconds to hours).

In addition, to increase the memory-capacity of such a system by introducing soft bounds for the synaptic weights, it is necessary to provide a mechanism that can realize the desired state dependence in the synaptic weight-update transfer function[45]. This can be achieved by engineering the conductance change properties of the single memristive device, or by designing hybrid memristive-CMOS neuromorphic circuits interfaced with one or more memristive devices[11,79]. Alternatively, one can use multiple binary memristive devices with probabilistic switching in combination with an analog circuit designed to properly control their switching probability.

As evident from the list of ingredients and recipe provided, it is now possible to build ultra low power massively parallel arrays of processing elements that implement "beyond-von Neumann", "in-memory computing" mixed signal hybrid memristive-CMOS neural processing systems.

It is important to realize that for data-intense processing applications these neuromorphic systems should be used to complement, rather than replace, traditional von Neumann architectures. They could be considered as the *cherry on the cake* of a complex AI inference engine, that enables always-on neural processing, with life-long learning abilities. In this scenario, the hybrid memristive-CMOS neuromorphic computing system would carry out low-power computation acting as alow accuracy predictive "watch-dog" to quickly activate more powerful von Neumann architectures for high accuracy recognition, as soon as events of interest are detected.

On the other hand, there are many applications where these hybrid neuromorphic systems would represent both the cherry *and* the cake together: these are IoT, edge-computing, and perception-action tasks that are solved efficiently by biological systems but have been proven to be "difficult" for artificial intelligence algorithms[80]. This difficulty could be measured with different performance metrics that could range from the physical size and energy consumption requirements to latency, adaptation, and ability to learn in continuous time closed-loop setups. By appropriately mixing all the ingredients and integrating them with mixed-signal analog/digital neuromorphic systems, it will be possible to produce computing systems that can directly *emulate* their biological counterparts. This emulation feature, which derives from the exploitation of the physics of the new materials and memory technologies being developed, is the key element for building efficient computing devices that can interact with the environment to solve artificial intelligence tasks in the real physical world, rather than simulating these solutions with general purpose computers. In other words, it is not very useful to simulate the bee brain on a supercomputer because it will never fly.

ACKNOWLEDGMENTS

We wish to acknowledge Melika Payvand and Regina Dittmann for the constructive comments on the manuscript. The illustration of Fig. 1 was kindly provided by University of Zurich, Information Technology, MELS/SIVIC, Sarah Steinbacher. This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 724295 (NeuroAgents).

[1] E. O. Neftci, "Data and power efficient intelligence with neuromorphic learning machines," iScience **5**, 52 – 68 (2018).
[2] C. S. Thakur, J. L. Molin, G. Cauwenberghs, G. Indiveri, K. Kumar, N. Qiao, J. Schemmel, R. Wang, E. Chicca, J. Olson Hasler, J.-s. Seo, S. Yu, Y. Cao, A. van Schaik, and R. Etienne-Cummings, "Large-scale neuromorphic spiking array processors: A quest to mimic the brain," Frontiers in Neuroscience **12**, 891 (2018).
[3] Y. Li, Z. Wang, R. Midya, Q. Xia, and J. J. Yang, "Review of memristor devices in neuromorphic computing: materials sciences and device challenges," Journal of Physics D: Applied Physics **51**, 503002 (2018).
[4] Y. van De Burgt, A. Melianas, S. T. Keene, G. Malliaras, and A. Salleo, "Organic electronics for neuromorphic computing," Nature Electronics **1**, 386–397 (2018).
[5] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, *et al.*, "Neuromorphic computing using non-volatile memory," Advances in Physics: X **2**, 89–124 (2017).
[6] C. Mead, "Neuromorphic electronic systems," Proceedings of the IEEE **78**, 1629–36 (1990).
[7] S. Furber, F. Galluppi, S. Temple, and L. Plana, "The SpiNNaker project," Proceedings of the IEEE **102**, 652–665 (2014).
[8] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, *et al.*, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **34**, 1537–1557 (2015).
[9] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro **38**, 82–99 (2018).
[10] D. Ielmini and R. Waser, *Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications* (John Wiley & Sons, 2015).
[11] I. Boybat, M. L. Gallo, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, E. Eleftheriou, *et al.*, "Neuromorphic computing with multi-memristive synapses," Nature communications **9**, 2514 (2018).
[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).
[13] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks **61**, 85–117 (2015).
[14] A. Sebastian, M. L. Gallo, and E. Eleftheriou, "Computational phase-change memory: beyond von neumann computing," Journal of Physics D: Applied Physics **52**, 443002 (2019).
[15] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, "Equivalent-accuracy accelerated neural-network training using analogue memory," Nature **558**, 60–67 (2018).
[16] S. Dai, Y. Zhao, Y. Wang, J. Zhang, L. Fang, S. Jin, Y. Shao, and J. Huang, "Recent advances in transistor-based artificial synapses," Advanced Functional Materials **0**, 1903700 (2019), https://onlinelibrary.wiley.com/doi/pdf/10.1002/adfm.201903700.

[17] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, "Analog memristive synapse in spiking networks implementing unsupervised learning," Frontiers in neuroscience **10**, 1–13 (2016).
[18] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," Nano letters **10**, 1297–1301 (2010).
[19] J. J. Yang and Q. Xia, "Organic electronics: Battery-like artificial synapses," Nature materials **16**, 396 (2017).
[20] R. Berdan, E. Vasilaki, A. Khiat, G. Indiveri, A. Serb, and T. Prodromakis, "Emulating short-term synaptic dynamics with memristive devices," Scientific Reports **6**, 1–9 (2016).
[21] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, "Neuromorphic electronic circuits for building autonomous cognitive systems," Proceedings of the IEEE **102**, 1367–1388 (2014).
[22] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems," Proceedings of the IEEE **103**, 1379–1397 (2015).
[23] G. Indiveri and Y. Sandamirskaya, "The importance of space and time for signal processing in neuromorphic agents," IEEE Signal Processing Magazine **36**, 16–28 (2019).
[24] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science **345**, 668–673 (2014).
[25] J. Backus, "Can programming be liberated from the von neumann style?: a functional style and its algebra of programs," Communications of the ACM **21**, 613–641 (1978).
[26] G. Indiveri, B. Linares-Barranco, R. Legenstein, G. Deligeorgis, and T. Prodromakis, "Integration of nanoscale memristor synapses in neuromorphic computing architectures," Nanotechnology **24**, 384010 (2013).
[27] J. Von Neumann, "First draft of a report on the edvac," IEEE Annals of the History of Computing **15**, 27–75 (1993).
[28] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull. Math. Biophys. **5**, 115–133 (1943).
[29] M. Shadlen and J. Movshon, "Synchrony unbound: a critical evaluation of the temporal binding hypothesis," Neuron **24**, 67–77 (1999).
[30] K. Boahen, "Communicating neuronal ensembles between neuromorphic chips," in *Neuromorphic Systems Engineering*, edited by T. Lande (Kluwer Academic, Norwell, MA, 1998) pp. 229–259.
[31] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, "A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs)," Biomedical Circuits and Systems, IEEE Transactions on **12**, 106–122 (2018).
[32] N. Qiao, C. Bartolozzi, and G. Indiveri, "An ultralow leakage synaptic scaling homeostatic plasticity circuit with configurable time scales up to 100 ks," IEEE Transactions on Biomedical Circuits and Systems (2017), 10.1109/TBCAS.2017.2754383.
[33] X. Zhang, S. Liu, X. Zhao, F. Wu, Q. Wu, W. Wang, R. Cao, Y. Fang, H. Lv, S. Long, Q. Liu, and M. Liu, "Emulating short-term and long-term plasticity of bio-synapse based on cu/a-si/pt memristor," IEEE Electron Device Letters **38**, 1208–1211 (2017).
[34] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. Gimzewski, and M. Aono, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses," Nature Materials **10**, 591–595 (2011).
[35] T. Werner, E. Vianello, O. Bichler, A. Grossi, E. Nowak, J.-F. Nodin, B. Yvert, B. D. Salvo, and L. Perniola, "Experimental demonstration of short and long term synaptic plasticity using oxRAM multi k-bit arrays for reliable detection in highly noisy input data," in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016) pp. 16–6.
[36] J. Yoon, H. Jung, Z. Wang, K. M. Kim, H. Wu, V. Ravichandran, Q. Xia, C. S. Hwang, and J. J. Yang, "An artificial nociceptor based on a diffusive memristor," Nature communications **9**, 417 (2018).
[37] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Q. Wu, M. Barnell, G.-L. Li, H. L. Xin, R. S. Williams, Q. Xia, and J. J. Yang, "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing," Nature materials **16**, 101 (2017).
[38] J. Xiong, R. Yang, J. Shaibo, H. M. Huang, H. K. He, W. Zhou, and X. Guo, "Bienenstock, cooper, and munro learning rules realized in second-order memristors with tunable forgetting rate," Advanced Functional Materials **29**, 1807316 (2019).
[39] M. Payvand, M. V. Nair, L. K. Müller, and G. Indiveri, "A neuromorphic systems approach to in-memory computing with non-ideal memristive devices: From mitigation to exploitation," Faraday Discussions **213**, 487–510 (2019).
[40] S. Pi, C. Li, H. Jiang, W. Xia, H. Xin, J. J. Yang, and Q. Xia, "Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension," Nature nanotechnology **14**, 35 (2019).
[41] Q. Xia and J. J. Yang, "Memristive crossbar arrays for brain-inspired computing," Nature materials **18**, 309 (2019).
[42] M. V. Nair and G. Indiveri, "A differential memristive current-mode circuit," European patent application EP 17183461.7 (2017), filed 27.07.2017.
[43] M. Payvand, L. K. Muller, and G. Indiveri, "Event-based circuits for controlling stochastic learning with memristive devices in neuromorphic architectures," in *Circuits and Systems (ISCAS), 2018 IEEE International Symposium on* (IEEE, 2018) pp. 1–5.
[44] E. O. Neftci, B. U. Pedroni, S. Joshi, M. Al-Shedivat, and G. Cauwenberghs, "Stochastic synapses enable efficient brain-inspired learning machines," Frontiers in Neuroscience **10**, 241 (2016).
[45] S. Brivio, D. Conti, M. V. Nair, J. Frascaroli, E. Covi, C. Ricciardi, G. Indiveri, and S. Spiga, "Extended memory lifetime in spiking neural networks employing memristive synapses with nonlinear conductance dynamics," Nanotechnology **30**, 015102 (2019).
[46] N. Diederich, T. Bartsch, H. Kohlstedt, and M. Ziegler, "A memristive plasticity model of voltage-based stdp suitable for recurrent bidirectional neural networks in the hippocampus," Scientific Reports (Nature Publisher Group) **8**, 1–12 (2018).
[47] A. Fantini, L. Goux, R. Degraeve, D. J. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in hfo 2 RRAM," in *2013 5th IEEE International Memory Workshop* (IEEE, 2013) pp. 30–33.
[48] M. Suri and V. Parmar, "Exploiting intrinsic variability of filamentary resistive memory for extreme learning machine architectures," IEEE transactions on nanotechnology **14**, 963–968 (2015).
[49] A. Schönhals, R. Waser, and D. J. Wouters, "Improvement of SET variability in TaOxbased resistive RAM devices," Nanotechnology **28**, 465203 (2017).
[50] A. Prakash, D. Deleruyelle, J. Song, M. Bocquet, and H. Hwang, "Resistance controllability and variability improvement in a taox-based resistive memory for multilevel storage application," Applied Physics Letters **106**, 233104 (2015).
[51] B. Govoreanu, D. Crotti, S. Subhechha, L. Zhang, Y. Y. Chen, S. Clima, V. Paraschiv, H. Hody, C. Adelmann, M. Popovici, O. Richard, and M. Jurczak, "A-VMCO: A novel forming-free, self-rectifying, analog memory cell with low-current operation, nonfilamentary switching and excellent variability," in *2015 Symposium on VLSI Technology (VLSI Technology)* (IEEE, 2015) pp. T132–T133.
[52] X. Sheng, C. E. Graves, S. Kumar, X. Li, B. Buchanan, L. Zheng, S. Lam, C. Li, and J. P. Strachan, "Low-conductance and multilevel CMOS-integrated nanoscale oxide memristors," Advanced Electronic Materials , 1800876 (2019).
[53] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of computer and system sciences **55**, 119–139 (1997).
[54] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," Science **304**, 78–80 (2004).
[55] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," Neural Computation **14**, 2531–2560 (2002).
[56] S. Sheik, M. Coath, G. Indiveri, S. Denham, T. Wennekers, and E. Chicca, "Emergent auditory feature tuning in a real-time neuromorphic VLSI system," Frontiers in Neuroscience **6** (2012), 10.3389/fnins.2012.00017.
[57] O. Richter, R. F. Reinhard, S. Nease, J. Steil, and E. Chicca, "Device mismatch in a neuromorphic system implements random features for regression," in *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (IEEE, 2015) pp. 1–4.
[58] A. Das, P. Pradhapan, W. Groenendaal, P. Adiraju, R. T. Rajan, F. Catthoor, S. Schaafsma, J. L. Krichmar, N. D. Dutt, and C. V. Hoof, "Unsupervised heart-rate estimation in wearables with liquid states and a probabilis-

tic readout," Neural networks **99**, 134–147 (2018).
[59]E. Donati, M. Payvand, N. Risi, R. Krause, K. Burelo, T. Dalgaty, E. Vianello, and G. Indiveri, "Processing EMG signals using reservoir computing on an event-based neuromorphic system," in *Biomedical Circuits and Systems Conference, (BioCAS), 2018* (IEEE, 2018) pp. 1–4.
[60]F. Bauer, D. Muir, and G. Indiveri, "Real-time ultra-low power ecg anomaly detection using an event-driven neuromorphic processor," Biomedical Circuits and Systems, IEEE Transactions on (2019), (in press).
[61]S. Gaba, P. Sheridan, J. Zhou, S. Choi, and W. Lu, "Stochastic memristive devices for computing and neuromorphic applications," Nanoscale **5**, 5872–5878 (2013).
[62]S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z.-Q. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide RRAM," IEEE Transactions on Electron Devices **63**, 1508–1515 (2016).
[63]J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing," Nature nanotechnology **8**, 13–24 (2013).
[64]M. Al-Shedivat, R. Naous, G. Cauwenberghs, and K. N. Salama, "Memristors empower spiking neurons with stochasticity," IEEE Journal on Emerging and Selected Topics in Circuits and Systems **5**, 242–253 (2015).
[65]M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Bio-inspired stochastic computing using binary CBRAM synapses," Electron Devices, IEEE Transactions on **60**, 2402–2409 (2013).
[66]S. Balatti, S. Ambrogio, R. Carboni, V. Milo, Z. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Physical unbiased generation of random numbers with coupled resistive switching devices," IEEE Transactions on Electron Devices **63**, 2029–2035 (2016).
[67]S. Habenschuss, Z. Jonke, and W. Maass, "Stochastic computations in cortical microcircuit models," PLoS computational biology **9**, e1003311 (2013).
[68]A. Destexhe and D. Contreras, "Neuronal computations with stochastic network states," Science **314**, 85–90 (2006).
[69]S. Fusi and W. Senn, "Eluding oblivion with smart stochastic selection of synaptic updates," Chaos, An Interdisciplinary Journal of Nonlinear Science **16**, 1–11 (2006).
[70]I. Ginzburg and H. Sompolinsky, "Theory of correlations in stochastic neural networks," Physical Review E **50**, 3171–91 (1994).
[71]T. Dalgaty, M. Payvand, B. De Salvo, J. Casaz, G. Lama, E. Nowak, G. Indiveri, and E. Vianello, "Hybrid cmos-rram neurons with intrinsic plasticit," in *International Symposium on Circuits and Systems (ISCAS), 2019* (IEEE, 2019).
[72]A. Yousefzadeh, E. Stromatias, M. Soto, T. Serrano-Gotarredona, and B. Linares-Barranco, "On practical issues for stochastic stdp hardware with 1-bit synaptic weights," Frontiers in neuroscience **12** (2018).
[73]J. Leugering and G. Pipa, "A unifying framework of synaptic and intrinsic plasticity in neural populations," Neural computation **30**, 945–986 (2018).
[74]S. Fusi and L. Abbott, "Limits on the memory storage capacity of bounded synapses," Nature Neuroscience **10**, 485–493 (2007).
[75]S. Ganguli, D. Huh, and H. Sompolinsky, "Memory traces in dynamical systems," Proceedings of the National Academy of Sciences **105**, 18970–18975 (2008).
[76]J. Frascaroli, S. Brivio, E. Covi, and S. Spiga, "Evidence of soft bound behaviour in analogue memristive devices for neuromorphic computing," Scientific reports **8**, 71–78 (2018).
[77]P. Del Giudice, S. Fusi, and M. Mattia, "Modeling the formation of working memory with networks of integrate-and-fire neurons connected by plastic synapses," Journal of Physiology Paris 97 , 659–681 (2003).
[78]F. Zenke, E. J. Agnes, and W. Gerstner, "Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks," Nature Communications **6**, 1–13 (2015).
[79]J. Bill and R. Legenstein, "A compound memristive synapse model for statistical learning through stdp in spiking neural networks," Frontiers in neuroscience **8**, 1–18 (2014).
[80]G. Plastiras, M. Terzi, C. Kyrkou, and T. Theocharidcs, "Edge intelligence: Challenges and opportunities of near-sensor machine learning applications," in *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)* (IEEE, 2018) pp. 1–7.